ADVISE ONLY”

Alchieri Leonardo - Badalotti Davide - Bonardi Pietro - Boschi Giulia

Università degli studi
Milano-Bicocca

Outline

- Esplorazione
- Obbiettivi
- Feature Selection
- Clustering
- Risultati

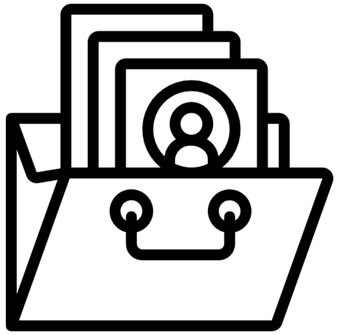
DATASET DESCRIPTION

Struttura

- 5000 records
- 17 attributi

Protocollo MIFID II

- Esperienza finanziaria
- Utilità per il consulente



Dataset Description

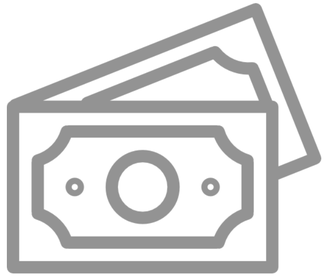
Struttura delle variabili

Legate alla persona

Sex, Age, Prov,
RiskPropension, PanicMood,
InheritanceIndex, …

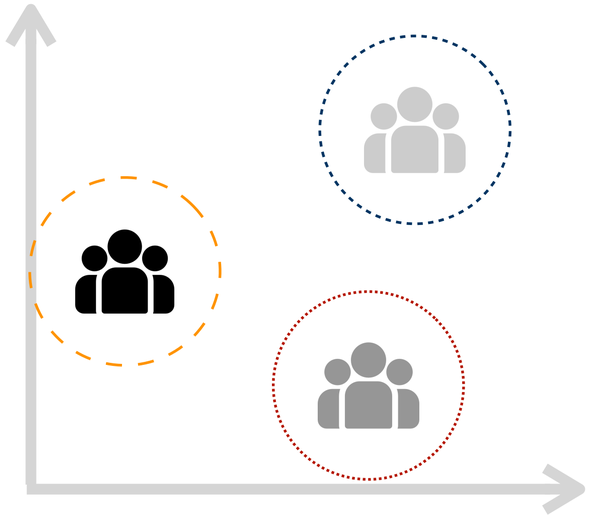
Legata al portafoglio

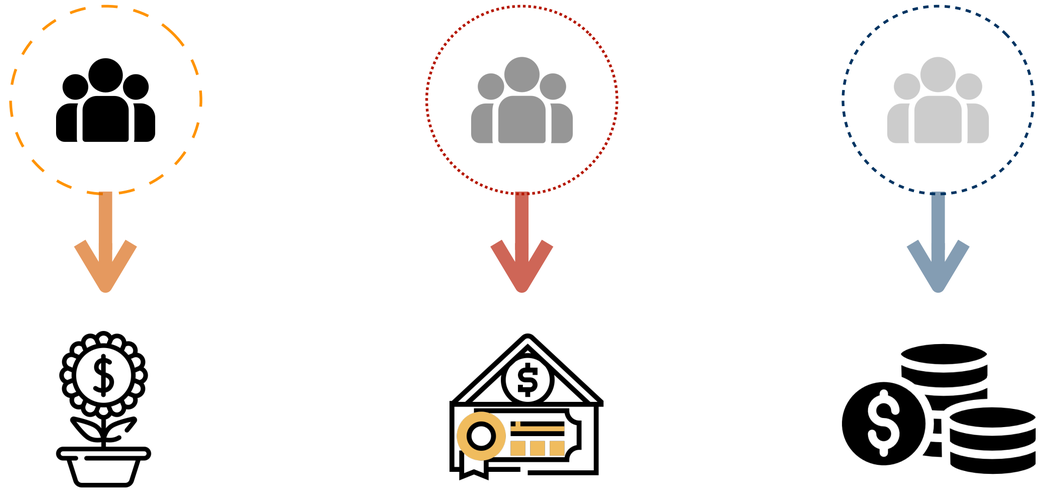
PortfolioRisk, AuM,
BondInvestments, Cash
EquityInvestments,…



OBBIETTIVI

- Uso di variabili persona per fare clustering
- Confronto tra variabili personali e di portafoglio
- Consiglio di prodotti finanziari ad ogni gruppo





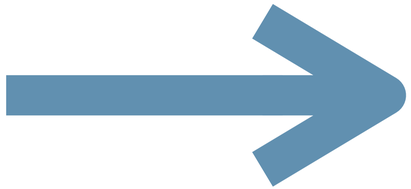
CONSIDERAZIONI

Non è possibile utilizzare tutte le variabili del dataset:

- Sparsificazione
- Difficile interpretazione

Impossibile approccio Brute Force:
2^{14} combinazioni possibili

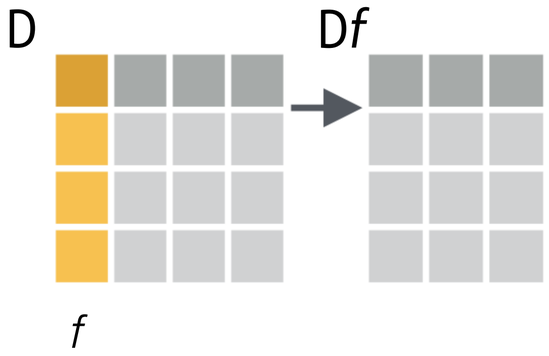
No riduzione dimensionale:
Attributi difficilmente interpretabili



Feature Ranking

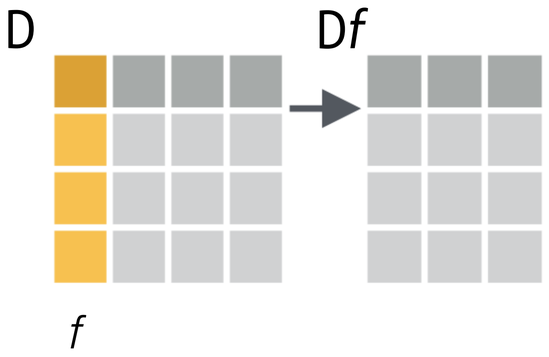
SRANK ALGORITHM

- Classifica l'**importanza** di ciascuna feature per un processo di clustering.
- Basato sul calcolo dell'**entropia** del dataset.



Maggiore è l'entropia di D*f*, maggiore sarà l'importanza di *f* nel dataset.

SRANK ALGORITHM



1. Matrice delle distanze: *D*

$$D_{ij} = \sqrt{\sum_k (x_{ik}^2 - x_{jk}^2)}$$

2. Matrice delle similarità: *S*

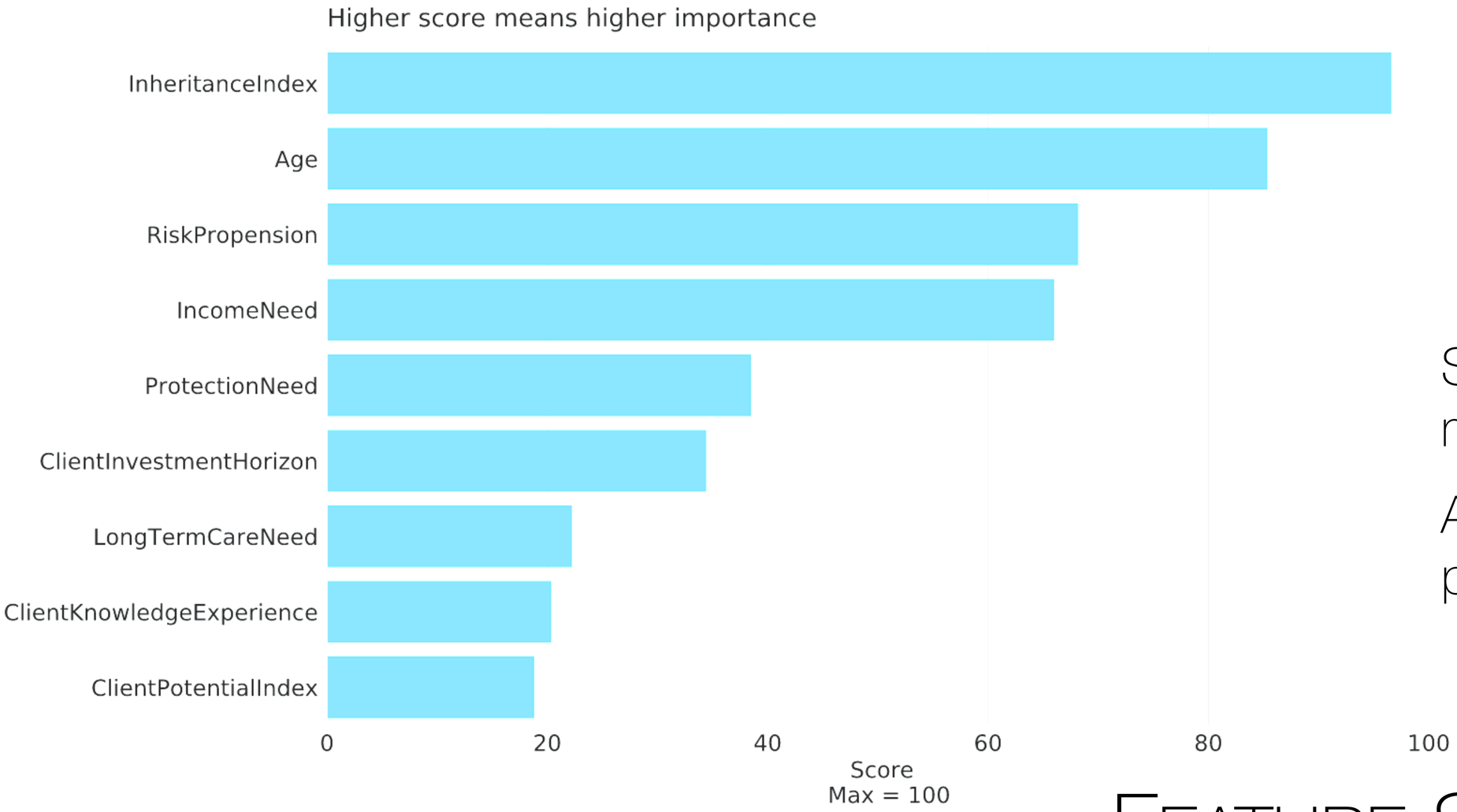
$$S_{ij} = e^{-\alpha \cdot D_{ij}}$$

3. Entropia del dataset: E*f*

$$Ef = -\sum_{ij} (S_{ij} \cdot \log S_{ij}) + (1 - S_{ij}) \cdot \log (1 - S_{ij})$$

Maggiore E*f*, maggiore importanza di f

Feature Score



Si valuta il subset migliore:

Aggiunta variabili a partire dalla cima

INTRODUZIONE

Diversi approcci basati su:

- Prototipo, e.g. K medie
- Densità, e.g. DBSCAN o HDBSCAN
- Probabilità, e.g. Gaussian Mixture

→ Alcune problematiche

Trial and error per determinare la scelta del modello e delle variabili, con supporto di visualizzazioni parziali.

Variabili scelte empiricamente da quelle determinate tramite Feature Selection.

HDBSCAN

Con variabili

- Età
- Necessità di reddito da cedole/ dividendi
- Propensione a prendere rischi
- Protezione da perdite finanziare
- Bisogno di ottimizzazione del patrimonio per gli eredi

Hierarchical Density-Based Spatial Clustering of Applications with Noise

- Trova cluster locali usando un processo di vicinanza tra punti (in distanza euclidea)
- Raggruppa cluster locali in una struttura gerarchica
- Individua outlier, o **noise**.

VALIDAZIONE

Possibili metodi

Parametrici ← → Grafici

↓

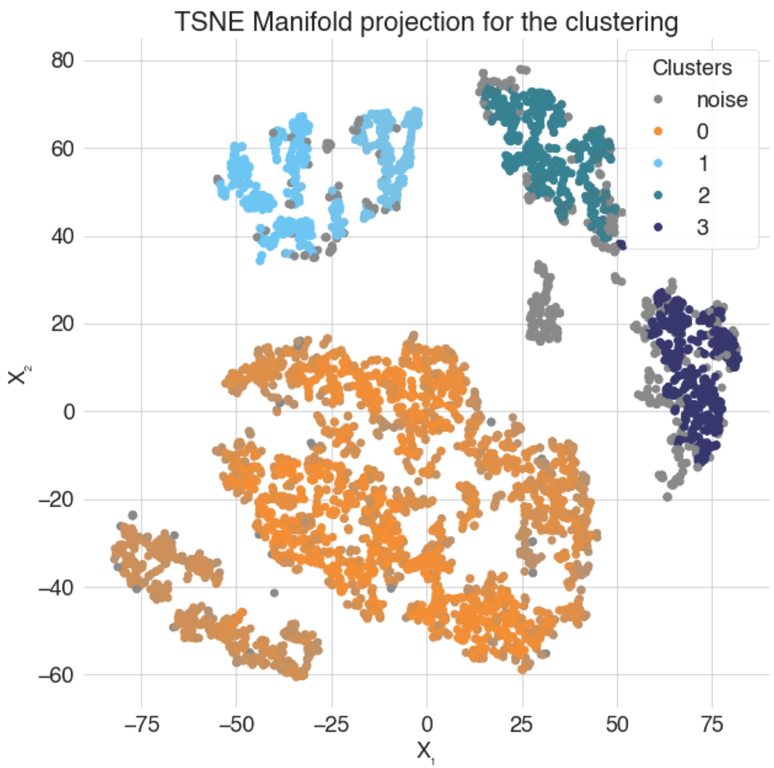
Nostra proposta:

Riduzioni dimensionale come validazione: Manifold Learning tramite **TSNE**

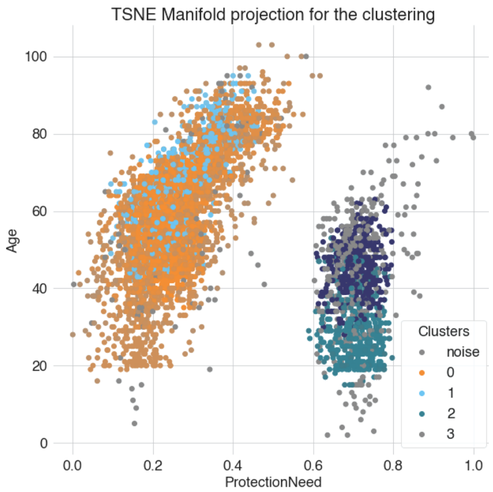
TSNE trova la pairwise probabiliy stimate lungo una t di Student di tutti i punti per costruire delle nuove variabili. → molto buono a distinguere pattern locali

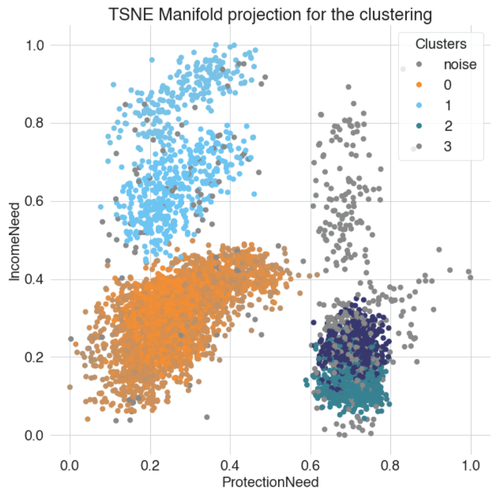
VALIDAZIONE

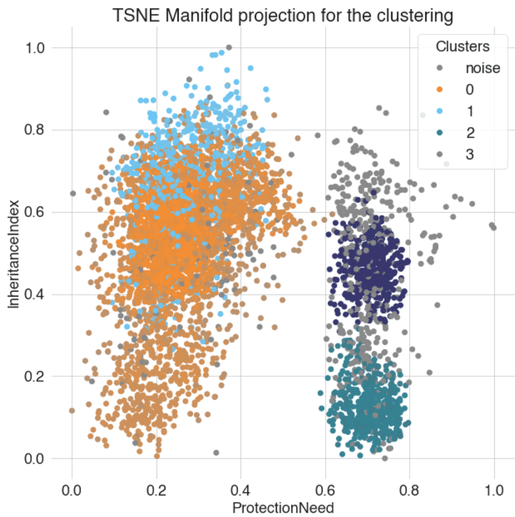
Validazione con TSNE



Proiezione su due assi







VARIABILI DI CLUSTER

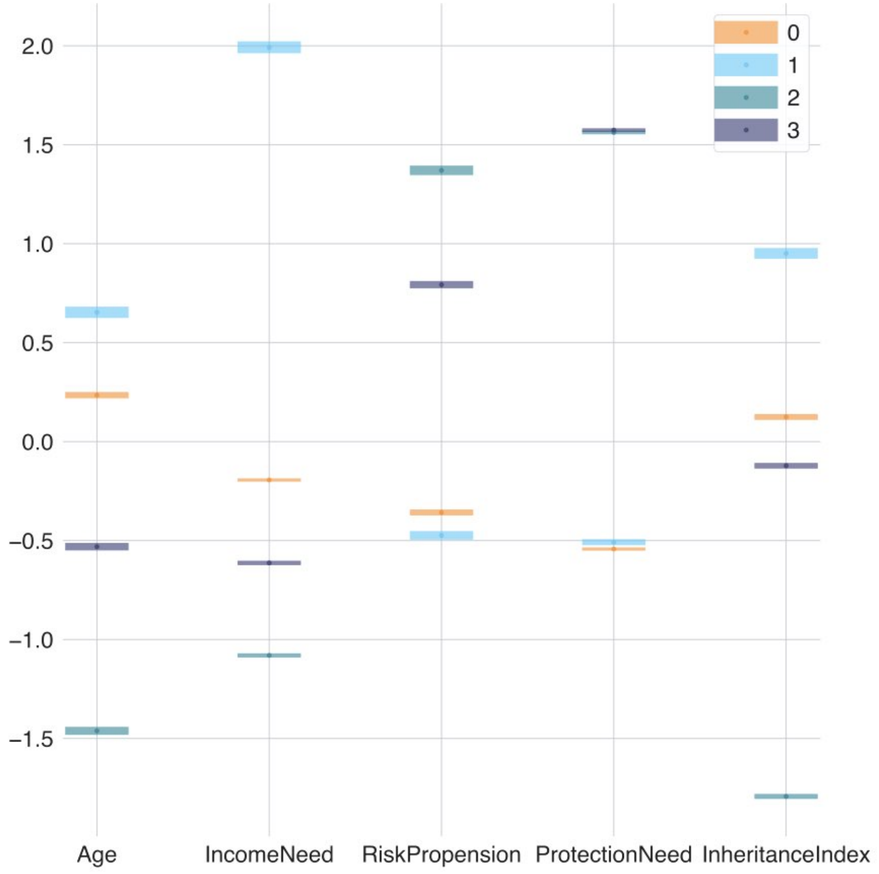
Le variabili sono significativamente distinte per ogni cluster

Cluster 2 → Young
Alta propensione al rischio finanziario, ma bisogno prodotti sicuri;
Bassa necessità di reddito da cedole/dividendi.

Cluster 3 → Adult
Simile al gruppo Young, dal quale si distacca principalmente per i valori di Inheritance Index.

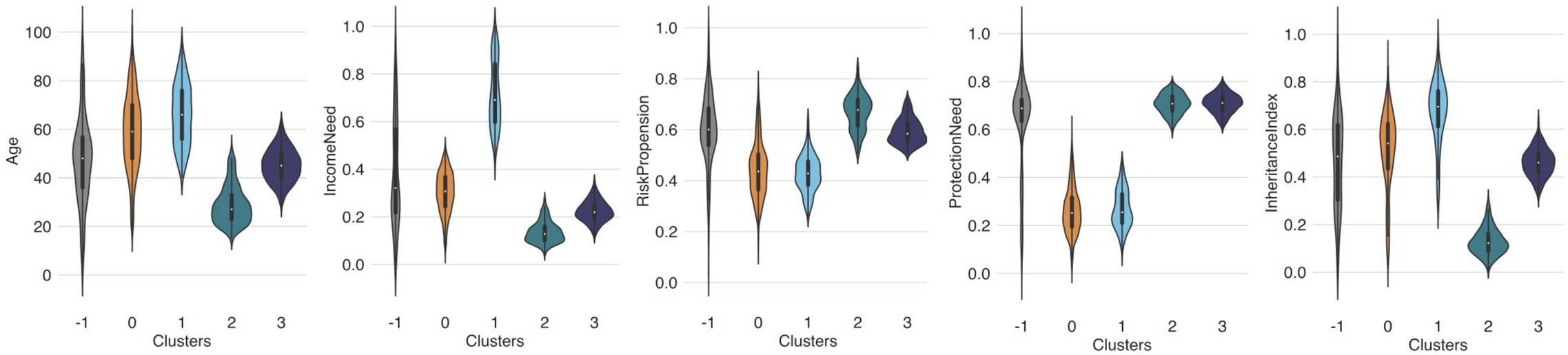
Cluster 0 → Adult/mature
Income need simile ai gruppi più giovani, ma ha una minor propensione al rischio e un minor bisogno di prodotti sicuri.

Cluster 1 → Mature
Distacca gli altri gruppi per necessità di reddito da cedole/dividendi, ma ha una bassa capacità di sopportare rischi finanziari; bisogno di ottimizzazione successoria/fiscale.



Intervalli di confidenza della media di ogni cluster per le diverse variabili.

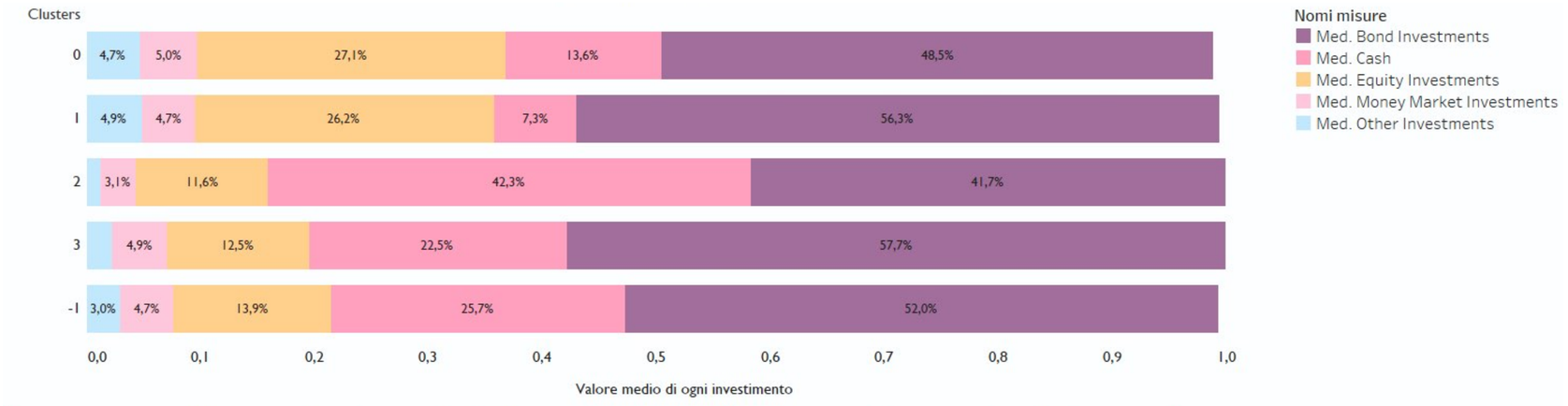
VARIABILI DI CLUSTER



Distribuzione delle variabili all'interno di ogni cluster.

- I cluster 0 (**Adult/Mature**) e 1 (**Mature**) hanno distribuzioni più allungate rispetto a 2 (**Young**) e 3 (**Adult**) che sono maggiormente concentrati intorno alla media
- Income need e protection need sono variabili molto differenzianti
- I dati relativi al Cluster -1 (**Noise**) hanno distribuzione omogenea.

PORTAFOGLIO TIPO



Divisione percentuale media del portafoglio tipo di ogni cluster tra i diversi tipi di investimento

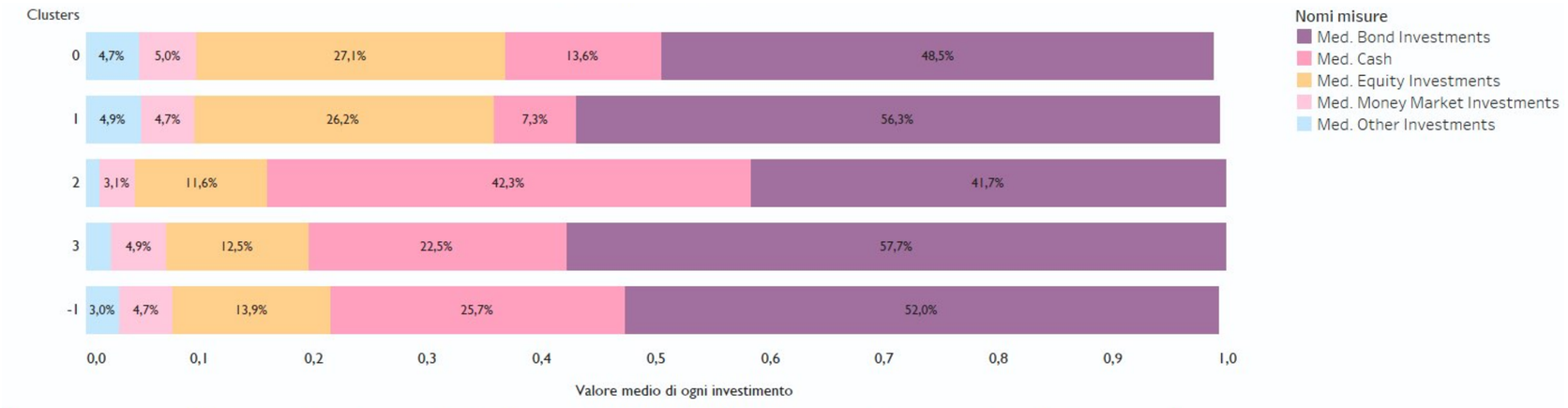
Cluster 2 → Young Maggior investitore in Cash, in linea con l'alto bisogno di protezione.

Cluster 3 → Adult Maggior investitore in Bond.

Cluster 0 → Adult/mature Maggior investitore in Azioni, in linea con il basso bisogno di prodotti sicuri.

Cluster 1 → Mature Minor investitore in Cash, in linea con l'alto necessità di reddito da cedole.

PRODOTTI MIGLIORI



Cluster 2 → Young Bassa disponibilità finanziaria: maggior acquisto di Bond o Moneta per investimenti a lungo termine e sicuri oppure Azioni a basso rischio.

Cluster 3 → Adult Investimenti in linea con caratteristiche personali: aumentare acquisto di Moneta o Azioni

Cluster 0 → Adult/mature Ridurre gli investimenti in azioni ed aumentare Bond e Cash oppure in Moneta se preferisse minore liquidità e orizzonte temporale più breve.

Cluster 1 → Mature Aumentare la Liquidità per aumentare la capacità di sopportare rischi finanziari e per facilitare le operazioni di ottimizzazione successoria/fiscale.

Cluster -1 → Noise Soluzioni di tipo tailor made.

Sviluppi Futuri

- Analisi ulteriori con appoggio ad esperti di dominio
- Sfruttare strutture gerarchiche per sottocategorie

REFERENZE

I. Moulavi, Davoud, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. “Density-Based Clustering Validation.” In Proceedings of the 2014 SIAM International Conference on Data Mining, 839–847.

II. Rahmah, Nadia, and Imas Sukaesih Sitanggang. “Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra.” IOP Conference Series: Earth and Environmental Science 31 (January 2016): 012012.

III. Maaten, Laurens van der, and Geoffrey Hinton. “Visualizing Data Using T-SNE.” Journal of Machine Learning Research 9, no. Nov (2008): 2579–2605.

IV. Dash, M., & Liu, H. (2000). Feature selection for clustering. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 1805, pp. 110-121). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 1805). Springer Verlag.

V. Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander. “Density-Based Clustering Based on Hierarchical Density Estimates.” In *Advances in Knowledge Discovery and Data Mining*, edited by Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, 160–172. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013.

ALGORITMO SRANK

Per il calcolo di alpha:

$$S = e^{-\alpha \cdot D}$$
$$S = 0.5$$
$$0.5 = e^{-\alpha \cdot \overline{D}}$$

Un valore di S = 0.5 massimizza l'entropia

$$\alpha = -\frac{\ln 0.5}{\overline{D}}$$

TSNE

t-distributed **S**tochastic **N**eighbour-**E**mbedding.

- Valuta la probabilità di coppia che x_i e x_j siano nello stesso vicinato, con distanza Eucliudea. Si definisce affinché sia simmetrica come:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

- Le probabilità asimmetriche che solo x_i contenga x_j nel suo vicinato viene calcolata su una Gaussiana come:

$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

- Definendo le variabili in **basse dimensioni** come y_i e y_j, valuto la stessa probabilità per queste affinché sia simmetrica e usando una **distribuzione t di Student:**

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

- Si usa **Gradient Descent** per cercare l'ottimo della **divergenza di Kullback-Leibler**

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- Per determinare le varianze σ_i^2, si usa **binary search** tramite la **perplessità** (iperparametro):

$$-\sum_j p_{ij} \log_2 p_{ij} = \log_2(Perp(P_i))$$

APPENDICE B